W94–W98 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki402

PISCES: recent improvements to a PDB sequence culling server

Guoli Wang and Roland L. Dunbrack Jr*

Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

Received February 14, 2005; Revised and Accepted March 14, 2005

ABSTRACT

PISCES is a database server for producing lists of sequences from the Protein Data Bank (PDB) using a number of entry- and chain-specific criteria and mutual sequence identity. Our goal in culling the PDB is to provide the longest list possible of the highest resolution structures that fulfill the sequence identity and structural quality cut-offs. The new PISCES server uses a combination of PSI-BLAST and structure-based alignments to determine sequence identities. Structure alignment produces more complete alignments and therefore more accurate sequence identities than PSI-BLAST. PISCES now allows a user to cull the PDB by-entry in addition to the standard culling by individual chains. In this scenario, a list will contain only entries that do not have a chain that has a sequence identity to any chain in any other entry in the list over the sequence identity cut-off. PISCES also provides fully annotated sequences including gene name and species. The server allows a user to cull an input list of entries or chains, so that other criteria, such as function, can be used. Results from a search on the re-engineered RCSB's site for the PDB can be entered into the PISCES server by a single click, combining the powerful searching abilities of the PDB with PISCES's utilities for sequence culling. The server's data are updated weekly. The server is available at http://dunbrack.fccc.edu/pisces.

INTRODUCTION

For many purposes, it is useful to have a list of protein sequences of known structure from entries in the Protein Data Bank (PDB) (1) that fit certain user-defined criteria. These purposes include statistical analysis of certain structural features, such as side-chain rotamer distributions (2), benchmarking sequence alignment and structure prediction methods (3), and analysis of a set of proteins, such as those with a certain function (e.g. DNA-binding) (4). The criteria for limiting the list of sequences can be applied to each chain in a PDB file, to each entry as a whole or to the relationships between protein chains and entries. Chain-specific criteria include function, species, sequence length and whether the chain is presented with only Cα atoms in the PDB file or all atoms. Entry-specific criteria include the function of a complex, resolution, *R*-factor, experiment type (NMR or X-ray) and other structure quality criteria. Pairwise criteria include the sequence or structural similarity between proteins in the set.

There are a number of servers that provide such lists, including the REPRDB (5), ASTRAL (6), UniqueProt (7) sites and the PDB site itself. Each provides a variety of services for specific purposes. They vary in a number of respects, including how sequence identities are determined and the flexibility and extent of services available to the user. In this paper, we describe extensions to our PISCES server that provides some advantages over other such servers, depending on the purposes of the user.

First, we believe in general that local alignments are better than global alignments for the purpose of selecting sequences from the PDB dataset, since many proteins share homology in one domain that may represent only a portion of the complete sequence. Servers that use CLUSTAL W (8), which performs global alignments, may underestimate sequence identity for shared domains by aligning over unrelated portions of two sequences. It is also important that sequence alignments cover the full-length of related regions of each protein pair. BLAST [and to a lesser extent PSI-BLAST (9)] may overestimate sequence identity for distantly related proteins by providing incomplete alignments over only the most-conserved regions of the sequence pair. Overestimating sequence identities results in lists that are shorter than necessary since some pairs are removed because of the inaccurate estimation of sequence identity. Structure-based alignments are likely to provide the most complete and accurate alignments at low-sequence identity than any sequence-based alignment method (10). As described below, PISCES now uses a combination of

*To whom correspondence should be addressed. Tel: +1 215 728 2434; Fax: +1 215 728 2412; Email: Roland.Dunbrack@fccc.edu

structure alignments at low-sequence identity and sequence alignment using PSI-BLAST at high-sequence identity. Other criteria than strict sequence identity are sometimes used (7).

Second, sequence culling servers may provide services for a variety of purposes. [The verb to cull can be used in different senses. The original meaning (Webster's 1913) was to select or gather, as in flowers. More recently, the word is used in two ways: culling the weakest members of a herd, or culling the herd itself, leaving behind the healthiest animals. We are generally using it in the second sense, to cull the PDB by removing redundant or low-resolution structures.] In addition to culling the entire PDB, a user may wish to cull a subset of the PDB. For instance, a user may have a list of PDB entries that contain antibody structures and may want a subset of these with certain structural quality and mutual sequence identity. One can also apply the sequence identity 'by-entry' rather than 'by-chain'. We define this as removing a PDB entry (that may contain more than one chain) from the list if another entry already in the list has at least one chain that has a sequence identity, higher than the cut-off, to any chain in the proposed entry. This might be useful if one wants a list of unique complexes in the PDB. PISCES provides both these services.

Another aspect of such a server is the kind of annotation provided in the results. Each server returns a list of chains and some servers are able to return a list of FASTA-formatted sequences. Usually, the annotation is simply the chain ID and sometimes what the RCSB provides in its 'pdb_seqres.txt' file (ftp://ftp.rcsb.org/pub/pdb/derived_data/pdb_seqres.txt). However, the RCSB does not provide consistent annotation for each chain in each PDB file with multiple chains. For instance, PDB entry 2HLA has two chains. Chain A is the human histocompatibility antigen HLA-Aw68.1. Chain B is β2-microglobulin. In the FASTA-formatted sequence database from the PDB, these two chains are given as:

(i) 2hla_A mol:protein length:270 Human Class I Histocompatibility Antigen Aw
(ii) 2hla_B mol:protein length:99 Human Class I Histocompatibility Antigen Aw 6

In fact, the lines are unfortunately truncated, but even the annotation from the 'Sequence Details' resource for the entry 2HLA on RCSB's site has only this for Chain B: 'HUMAN CLASS I HISTOCOMPATIBILITY ANTIGEN AW 68.1'. There is, however, more information in the full mmCIF file, including the name 'beta-2-microglobulin' as well as the Swiss-Prot entry accession number from which additional information may be obtained.

Our PISCES server was developed from a previous version referred to as 'CulledPDB' begun in 1999. CulledPDB used BLAST to determine sequence identities, while PISCES has used PSI-BLAST to determine sequence identities (11) by building a position-specific scoring matrix or profile for each unique sequence in the PDB from a multi-round search of the non-redundant protein sequence database (12). PISCES also provides a resource for culling a user-input list of chains or entries from the PDB or from a set of non-PDB sequences provided as either a list of GenBank IDs or in the FASTA format. We have shown that the use of the PSI-BLAST profiles resulted in much longer lists than while using BLAST (11), while not exhibiting errors associated with using a global alignment algorithm, such as CLUSTAL W.

We have made several important improvements in PISCES that are described in this paper:

(i) We now use the structure alignment program CE (13) to refine the relationship determination between proteins and recalculate their sequence identities that are found by using PSI-BLAST.
(ii) PISCES now use CE-type sequence identity, i.e. identical pairs divided by all aligned pairs excluding gaps. PSI-BLAST, in contrast, calculates the sequence identity by the ratio of the number of identical pairs to the full-length of the alignment, including gaps. This change means that if two closely related proteins are aligned, but one has a large insertion, the sequence identity will remain high. This may occur for instance if a long disordered loop is engineered out of a protein to facilitate crystallization. The new definition of sequence identity is a more conservative criterion for sequence culling, since sequence identities are higher than they were previously using the PSI-BLAST values, and more sequences are removed for a target sequence identity.
(iii) PISCES is now able to cull by-entry in addition to by-chain, as described above.
(iv) PISCES now provides more annotation for each chain in the PDB than the PDB itself does. This annotation includes chain-specific gene names and functional and species information retrieved from the full mmCIF files of the PDB, and the Swiss-Prot and GenBank databases.
(v) From the re-engineered search engine of the PDB (now in beta release, http://pdbbeta.rcsb.org/pdb) (14), with a single click on the External Sites→PISCES menu on the PDB's search result page, all the hits or a selected subset of the results of a search can be transmitted to the PISCES site for culling using a user's input criteria for structural quality and sequence identity cut-offs.

METHODS

Database information and sequence annotation

PISCES uses the mmCIF-format files from RCSB to determine sequences, experiment type, resolution, *R*-factors and other features of PDB entries and chains. These mmCIF files are a result of the Uniformity Project, an effort by the RCSB to standardize and correct information across all the PDB files (15,16). Some missing values for resolution and *R*-factors are obtained from the PDBFINDER database (17). The PDB data used by the server are updated weekly.

PISCES works from a FASTA-formatted database of all sequences in the PDB called *pdbaa*, which is distinct from NCBI's database of the PDB sequences with the same name (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/pdbaa.gz). Our *pdbaa* database is available to users who want a complete set of sequences in the PDB. It is also used to provide the sequences and annotations returned by PISCES for subsets of the PDB. *pdbaa* provides basic information on chain length, experiment type (X-ray, NMR, etc.), resolution, *R*-factor and free *R*-factor as appropriate. These pieces of information are useful if a user wants to use a structure as a template for homology modeling, and wants to choose the best template by searching the whole PDB.

In order to integrate as much useful information as possible into the output of PISCES, we have gathered annotations on each PDB sequence from a number of sources. The goal is to have gene names for each sequence, Swiss-Prot or other database identifiers and species information. These are obtained from the following sources: the mmCIF files themselves, Swiss-Prot's Index of PDB entries (PDBTOSP.TXT, http://us.expasy.org/cgi-bin/lists?pdbtosp.txt), the Protein Identification Resource (PIR) and NCBI's non-redundant protein sequence database, in that order to obtain the desired information.

Sequence identities

Our goal in culling the PDB is to provide the longest list possible of the highest resolution structures that fulfill the sequence identity and structural quality cut-offs. For this purpose, we need to identify potential evolutionary relationships and complete sequence alignment between related regions of each identified pair. Historically, 98% of requests to PISCES have been for culled lists at 25% identity or higher. We have determined that PSI-BLAST identifies 99.9% of such relationships with *E*-value of 1.0 or better (data not shown), and thus we do not use structure alignment to identify evolutionary relationships. Rather, we use the structure alignment program CE (13) to align structure pairs which PSI-BLAST determines as having a sequence identity of 50% or less or for which the PSI-BLAST alignment covers <80% of the shorter sequence in each pair. PSI-BLAST is used as described previously (11).

We have found some cases where structure alignment sequence identities are much lower than those calculated with PSI-BLAST even though the alignment lengths are comparable. This may occur when one single-domain protein is homologous to each of the two domains in another protein. CE may align the first protein to either of the two domains of the second protein, but not necessarily the more closely related one. To account for this, we use the sequence identity obtained by either PSI-BLAST or CE, whichever has the larger number of identically aligned residue pairs.

Culling by-chain and by-entry

With the sequence identities in hand, PISCES uses the method of Hobohm and Sander (18,19) to cull the sequences that pass the chain and entry criteria input by the user. The details have been described previously (11).

Culling can be performed on a chain-or entry-level basis. Culling by-chain means treating each chain in each PDB entry as a separate entity. This is the standard procedure for creating culled PDB sequence lists. Based on the requests from a number of users of PISCES, we have added another functionality to PISCES: culling 'by-entry'. For this procedure, the sequence identity between any two entries is defined as the highest sequence identity of any one chain in one entry with any one chain in the other entry. In this way, no two entries will appear in the same list if they share chains with sequence identity over the cut-off. PISCES further allows the user to choose whether to cull within each entry and allows the user to use another sequence identity cut-off for this culling procedure. So for instance, if an entry is a homodimer, the user can choose to have both sequences returned (no culling within entry) or just one of them (culling within entry by some value <100%).

Usage

PISCES first gives the user the option to: (i) cull the entire PDB; (ii) cull from a search at the PDB's re-engineered search engine site—this option takes the user to the PDB's web page (http://pdbbeta.rcsb.org) before returning to PISCES; (iii) cull from a user-input list of chains or entries; (iv) cull from a list of GenBank entries; and (v) cull from a FASTA-formatted set of sequences or from BLAST or PSI-BLAST output. For options (iv) and (v), sequence alignments are obtained by using PSI-BLAST on the list of input sequences to determine sequence identities. It is assumed that these sequences are not in the PDB and, therefore, structural criteria are not used.

Option (i) takes a user directly to a page for inputting structural quality criteria (experiment type, resolution, *R*-factor, Cα-status, chain lengths, etc.) and sequence identity cut-offs. Option (ii) take a user to the RCSB site or the user may start at the RCSB site directly. Once the RCSB server has returned a list of entries that satisfy search parameters, the user can click on the External Sites→PISCES menu on the RCSB page to return to PISCES and to a page displaying the list of hits (all or selected) from RCSB. Option (iii) takes a user to a page for inputting a list of PDB entries or chains. After the list of PDB entries or chains to be searched is confirmed by the user for Options (ii) and (iii), he/she is asked for the structural criteria to be used for culling. PISCES then confirms the input data and asks for user name, institution and email address. When the results are ready, almost always within minutes, the server sends an email to the user to download the results from a page given in the email. The results include a list of the input sequences (if used), the input cut-offs, the list of chains or entries that result from culling and a FASTA-formatted file of the selected sequences. These results are stored for 15 days.

In addition to the sequence culling service, PISCES also provides databases and programs that may prove useful to the user:

(i) A fully annotated *pdbaa* database. For each PDB chain, the description line includes chain length, resolution and *R*-value (for X-ray structures), protein descriptions, database name and entry name in the reference database, and species.

(ii) Two more condensed variants of *pdbaa*: *pdbaa.ent* and *pdbaa.nr*. *pdbaa.ent* has all redundant chains within entries removed and the removed chain IDs are placed at the end of the description line of the representative chains. *pdbaa.nr* is similar to *pdbaa.ent*, but it removes all redundant chains within the whole *pdbaa* instead of only within entries.

(iii) A standalone package for PISCES. Users can download this package and easily install it on local machines. The standalone version of PISCES has all the major features of the web-based PISCES server.

RESULTS AND DISCUSSION

We investigated the effects of using structure-based alignments on the results of PISCES. On the PDB dataset, the number of related pairs was: 637 115 according to BLAST (*E*-value better than 1.0, alignment length >30) and 758 411 according to PSI-BLAST (*E*-value better than 1.0, alignment length >30). In Figure 1, we have plotted sequence identities and alignment lengths for PSI-BLAST versus BLAST and

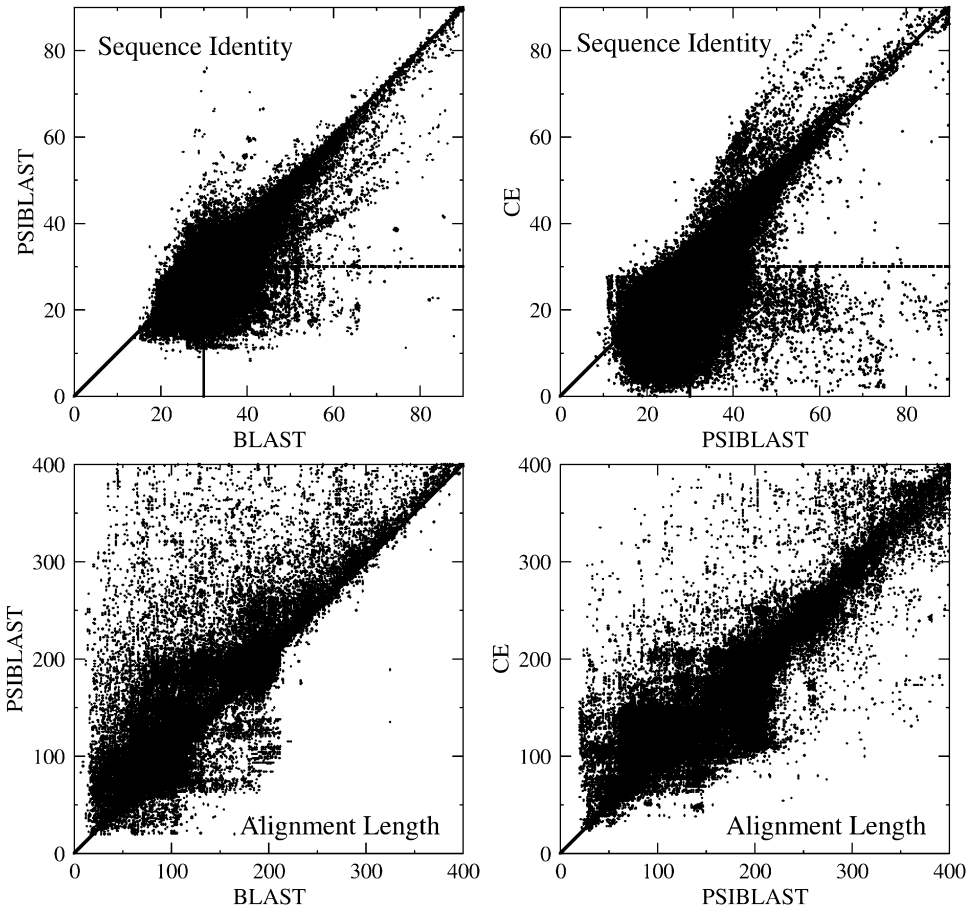

Figure 1. Scatterplots of sequence identities (top figures) and alignment lengths (bottom figures) for alignment pairs in common between PSI-BLAST and BLAST (left two figures) and CE and PSI-BLAST (right two figures). On the top two plots, a box is shown that contains points that would be used to eliminate sequences at 30% sequence identity using the method on the *x*-axis but not by the method on the *y*-axis.

CE versus PSI-BLAST. The bottom two plots show that the PSI-BLAST alignments are generally longer than the BLAST alignments, and the CE alignments are longer than the PSI-BLAST alignments. The top two plots show that the sequence identities are lower in PSI-BLAST than in BLAST and lower in CE than in PSI-BLAST. As noted earlier, this is due to the fact that incomplete alignments result in artificially higher sequence identities, because the alignment occurs over the most-conserved regions of the two sequences. We have boxed a region of the top two plots to show pairwise relationships that would be used to cull at 30% sequence identity. For instance, in the upper right figure these points would be used to cull using PSI-BLAST-determined sequence identities but not CE-determined identities, since the points in the box have PSI-BLAST sequence identity >30% but CE sequence identity <30%. There are far more points in this box than in the corresponding rectangular box for inclusion by PSI-BLAST but not by CE (to the left and above the green box).

In Table 1, we show the lengths of lists under various sequence identity cut-offs as determined by using BLAST, PSI-BLAST and the REPRDB server (5), which use CLUSTAL W and structure alignments. CE within PISCES provides longer lists at low-sequence identity than the sequence-based methods of BLAST and PSI-BLAST. At very high-sequence identity, the results are based primarily on PSI-BLAST calculations and thus are very similar. Actually, the difference between CE and PSI-BLAST is not only the list length, but also the content in the lists. For example, for lists with 25% sequence identity cut-off, the CE list has only 57 more chains than the PSI-BLAST list, but in fact the two lists have 157 different chains that are not in common. REPRDB produces very short lists at low-sequence identity for unknown reasons. Its lists are longer at high-sequence identity, probably because the global alignment algorithm produces alignments over

Table 1. Lengths of lists obtained using different sequence alignment methods

Percentage	BLAST	PSI-BLAST	REPRDB	CE
15	2092	1982	17	2041
20	2099	2268	75	2343
25	2184	2782	1492	2839
30	2562	3349	3677	3400
40	3863	4293	4570	4305
50	4878	5002	5200	5003
60	5408	5482	5695	5480
70	5845	5873	6102	5872
80	6194	6219	6566	6218
90	6686	6708	7462	6707

Criteria for inclusion in the lists: resolution ⩽3.0 Å; including Cα chains; excluding non-X-ray entries.

non-homologous regions of sequence pairs, thus lowering the sequence identity compared with a local alignment algorithm.

We now show an example of the improved annotations available in our *pdbaa* database. As shown in Introduction, chain-specific information is sometimes missing in the FASTA-formatted database available from the RCSB. For 2HLA, our *pdbaa* provides these annotations:

(i) 2HLAA 270 XRAY 2.60 0.173 NA CLASS I HISTOCOMPATIBILITY ANTIGEN (HLA-Aw68) <SWS 1A68_HUMAN> [HOMO SAPIENS]
(ii) 2HLAB 99 XRAY 2.60 0.173 NA BETA 2-MICROGLOBULIN <SWS B2MG_HUMAN> [HOMO SAPIENS]

After the chain identifier is the sequence length, experiment type, resolution, *R*-factor, free *R*-factor (not available in this case), the gene name, the database reference in angular brackets (in this case, Swiss-Prot entry B2MG_HUMAN) and the species in square brackets. While this information is available in the mmCIF file for 2HLA, it is not available in the FASTA sequence database provided by the RCSB.

Finally, we expect the link to PDB's search capabilities to be very useful. Since the establishment of this link, almost 10% of the requests to PISCES have come via the PDB search page. The re-engineered PDB page is only in beta test phase, and so we expect that PISCES will find additional utility when RCSB releases the final version of their search site.

ACKNOWLEDGEMENTS

We thank Dr Adrian A. Canutescu for advice on the web server implementation of PISCES. This work was funded by NIH Grants R01-HG02302 to R.L.D. and CA06972 to Fox Chase Cancer Center, as well as the Pennsylvania Tobacco Settlement and an appropriation from the Commonwealth of Pennsylvania. Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
2. Dunbrack,R.L.,Jr (2002) Rotamer libraries in the 21st century. *Curr. Opin. Struct. Biol*., **12**, 431–440.
3. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol*., **326**, 317–336.
4. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res*., **29**, 2860–2874.
5. Noguchi,T., Matsuda,H. and Akiyama,Y. (2001) PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res*., **29**, 219–220.
6. Chandonia,J.M., Hon,G., Walker,N.S., Lo Conte,L., Koehl,P., Levitt,M. and Brenner,S.E. (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res*., **32**, D189–D192.
7. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acids Res*., **31**, 3789–3791.
8. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
9. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of database programs. *Nucleic Acids Res*., **25**, 3389–3402.
10. Sauder,J.M., Arthur,J.W. and Dunbrack,R.L.,Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins*, **40**, 6–22.
11. Wang,G. and Dunbrack,R.L.,Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
12. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al*. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **32**, D35–D40.
13. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.
14. Bourne,P.E., Addess,K.J., Bluhm,W.F., Chen,L., Deshpande,N., Feng,Z., Fleri,W., Green,R., Merino-Ott,J.C., Townsend-Merino,W. *et al*. (2004) The distribution and query systems of the RCSB Protein Data Bank. *Nucleic Acids Res*., **32**, D223–D225.
15. Westbrook,J., Feng,Z., Jain,S., Bhat,T.N., Thanki,N., Ravichandran,V., Gilliland,G.L., Bluhm,W., Weissig,H., Greer,D.S. *et al*. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res*., **30**, 245–248.
16. Bhat,T.N., Bourne,P., Feng,Z., Gilliland,G., Jain,S., Ravichandran,V., Schneider,B., Schneider,K., Thanki,N., Weissig,H. *et al*. (2001) The PDB data uniformity project. *Nucleic Acids Res*., **29**, 214–218.
17. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
18. Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci*., **1**, 409–417.
19. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci*., **3**, 522–524.